DOCUMENT RESUME

ED 120 207 TM 005 166

AUTHOR Tinkelman, Sherman N.
TITLE Improving the Classroom Test: A Manual of Test Construction Procedures for the Classroom Teacher.
INSTITUTION New York State Education Dept., Albany. Bureau of Elementary and Secondary Educational Testing.
PUB DATE 75
NOTE 57p.

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage
DESCRIPTORS Elementary Secondary Education; Essay Tests; *Guidelines; Multiple Choice Tests; Objective Tests; Scoring; *Teachers; *Test Construction
IDENTIFIERS Completion Tests; Matching Tests; True False Tests

ABSTRACT

The purpose of this bulletin is to help the teacher derive the full value from the classroom test as an important educational tool. Principles of good test construction are explained and illustrated. Concrete examples of good procedure and poor procedure are offered. A checklist is presented of the salient points to keep in mind when preparing a test. No consideration is given in this bulletin to the curricular implications of classroom tests, vital as they are. The special concern of this bulletin is not "what to test", but rather "how to test." (RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

IMPROVING
the
CLASSROOM TEST

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED120207

A MANUAL of
TEST CONSTRUCTION PROCEDURES
for the
CLASSROOM TEACHER

TM005 166

The University of the State of New York
The State Education Department
Bureau of Elementary and Secondary Educational Testing
Albany

IMPROVING
the
CLASSROOM TEST

A MANUAL of
TEST CONSTRUCTION PROCEDURES
for the
CLASSROOM TEACHER

Reprinted 1975

The University of the State of New York
The State Education Department
Bureau of Elementary and Secondary Educational Testing
Albany, 1968

THE UNIVERSITY OF THE STATE OF NEW YORK

Regents of The University *(with years when terms expire)*

1984	JOSEPH W. MCGOVERN, A.B., J.D., L.H.D., LL.D., D.C.L., Litt.D., *Chancellor*	New York
1981	THEODORE M. BLACK, A.B., Litt.D., LL.D., Pd.D., *Vice Chancellor*	Sands Point
1978	ALEXANDER J. ALLAN, JR., LL.D., Litt.D.	Troy
1987	CARL H. PFORZHEIMER, JR., A.B., M.B.A., D.C.S., H.H.D.	Purchase
1975	EDWARD M. M. WARBURG, B.S., L.H.D.	New York
1980	JOSEPH T. KING, LL.B.	Shelter Island
1981	JOSEPH C. INDELICATO, M.D.	Brooklyn
1976	HELEN B. POWER, A.B., Litt.D., L.H.D., LL.D.	Rochester
1979	FRANCIS W. MCGINLEY, B.S., J.D., LL.D.	Glens Falls
1986	KENNETH B. CLARK, A.B., M.S., Ph.D., LL.D., L.H.D., D.Sc.	Hastings on Hudson
1983	HAROLD E. NEWCOMB, B.A.	Owego
1988	WILLARD A. GENRICH, LL.B., L.H.D.	Buffalo
1982	EMLYN I. GRIFFITH, A.B., J.D.	Rome
1977	GENEVIEVE S. KLEIN, B.S., M.A.	Bayside
1981	WILLIAM JOVANOVICH, A.B., LL.D., Litt.D., L.H.D.	Briarcliff Manor

President of The University and Commissioner of Education
EWALD B. NYQUIST

Executive Deputy Commissioner of Education
GORDON M. AMBACH

Contents

	PAGE
I. Planning the Classroom Test	7
II. Choosing an Item Type	8
A. Essay Questions	9
B. Completion Items	10
C. True-False Items	11
D. Multiple-Choice Items	13
E. Matching Items	14
III. Writing Effective Test Items	15
A. Essay Questions	15
B. Completion Items	19
C. True-False Items	22
D. Multiple-Choice Items	28
E. Matching Items	40
IV. The Mechanical Features of Classroom Tests	43
A. Item Format	43
B. Scoring Arrangements	44
C. Distribution of Correct Responses	45
D. Grouping and Arrangement of Items	46
E. Designating Credit Allowances	47
F. Directions for Answering Questions	47
G. Correction for Guessing	48
H. Allowing Choice of Items	49
I. Printing and Duplicating	50
V. Checklist for Reviewing Local School Tests	51
Bibliography	56

Preface

Important as Regents examinations and standardized tests are in the total school testing program, the backbone of day-by-day evaluation in the school is the classroom test.

The outstanding advantage of the classroom test is its flexibility. It can be used to stimulate pupil interest, to take inventory of pupil abilities prior to instruction, to measure progress during instruction, to establish readiness for new learning, to determine the need for, or the effectiveness of remedial work or to accomplish any other special purpose the teacher has in mind.

Its scope and emphasis can be tailormade to fit the needs of the class. The length of the test and the point in the learning program at which it is administered can be easily adjusted to the immediate classroom situation. Properly used, the classroom test can be extraordinarily helpful in guiding the educational development of the pupils and in measuring the effectiveness of instruction.

To provide these advantages, however, the classroom test must be well constructed. Using tests which provide erroneous information concerning the pupil is worse than not using tests at all. No teacher would care to purchase a standardized test with confusing directions or poorly framed items. Yet many teacher examinations exhibit flaws that seriously impair the validity and usefulness of the test scores. The high technical standards of standardized tests and Regents examinations should apply to classroom tests as well.

The purpose of this bulletin is to help the teacher derive the full value from the classroom test as an important educational tool. Principles of good test construction are explained and illustrated. Concrete examples of good procedure and poor procedure are offered. A checklist is presented of the salient points to keep in mind when preparing a test.

The bulletin does not attempt to provide a magic formula for making test writing a painless procedure. Constructing a good test is always a demanding task, challenging the best creative effort and professional judgment of the teacher. It is hoped, however, that a summary of current best thinking in the area of test construction, presented in

convenient form for ready reference, may prove useful to teachers interested in improving the quality of their classroom tests.

No consideration is given in this bulletin to the curricular implications of classroom tests, vital as they are. The special concern of this bulletin is not "what to test," but rather "how to test."

This manual was prepared by Sherman N. Tinkelman, Assistant Commissioner for Examinations and Scholarships.

I. Planning the Classroom Test

A Test as a Work Sample

A test is a work sample. It constitutes a selection of all the possible questions that might be asked. Pupil performance on this sample enables the teacher to generalize concerning progress in the total area from which the sample is drawn. Obviously, for such generalizations to be valid, it is essential that the test constitute a fair and representative sample of questions.

Before a single test question is written, therefore, the teacher should have clearly in mind what purpose the test is to serve, what skills and content areas are to be measured and what their relative weights should be. To ignore these factors at the outset in preparing a test is to court the danger that the test, if it constitutes a poor sample of questions, may provide a misleading picture of pupil progress.

Need for a Test Blueprint

In a test in American history, for example, some weighting of the political, social and economic aspects of history is inescapable. What should be the relative contribution of each to a single score purporting to indicate the pupil's general progress in all of American history? The problem cannot be solved by ignoring it. If the relative weights are not predetermined on the basis of the professional judgment of the teacher, they will be decided by the accidents of item selection. Depending upon the relative number of credits which may happen to be assigned, the test will automatically weight different aspects of history, different chronological periods, different types of skills and different historical topics.

This problem must be approached by the teacher in a direct and forthright manner. A good procedure is to outline beforehand, in written form, exactly what the test should cover if it is to serve its purpose well and exactly how much weight should be assigned to the various objectives and areas of the test in terms of their relative importance. The function of such a test outline or blueprint is to help assure a fair sampling of items to provide a test score that will be a valid index of overall pupil progress. A test drawn up in accordance with definite specifications is more likely to furnish a total score that is useful and meaningful than is a test whose items are assembled at random.

Flexibility of the Test Blueprint

There can be no single blueprint or type of blueprint that is best for all classroom tests. In a brief classroom quiz a rather simple listing of major areas and weights may be sufficient. The broad sweep of a final examination, however, may call for a comprehensive and detailed classification of course content and objectives. Each area of the school curriculum may have a different type of blueprint organization, depending upon its objectives. Each teacher may have a different distribution of emphasis. The same teacher may prepare different blueprints for different groups of pupils, depending upon the immediate purpose of the test in the classroom situation.

In any case, however, every test should be developed from a carefully drawn blueprint. If a test is to be worth the time and effort involved in preparing it and administering it, the scope and content of the test should be planned to assure that the test will achieve the purpose for which it is intended.

II. Choosing an Item Type

Relative Merits of Item Types

Are essay questions more reliable than objective-type questions? Are multiple-choice items better than true-false items? The relative merits of different types of test questions or items were long a popular topic for debate in educational circles. We now know that there is no general answer to such questions. There is no one type of item that is inherently superior to other types. How well an item is constructed is more important than the type of item it represents. Each type of item, moreover, has special strengths and weaknesses. The choice of item types for a classroom examination therefore devolves upon the special advantages and disadvantages of each for the purpose the teacher has in mind.

Classification of Item Types

Although there is a wide variety of different item forms in use today, it will not be necessary to consider in this bulletin more than the few basic item forms that account for the bulk of all items written. The principles leading to successful use of these basic forms can be applied generally to related forms.

Examination questions are generally classified as either essay or objective in type. While the dividing lines are sometimes difficult to draw, the term *essay* is usually applied to a question that allows the

pupil some degree of freedom to organize and express his answer in his own words. The answer required may range from a sentence to a dissertation, but the quality of the answer must be judged subjectively by a person who is informed in the subject. The *objective* type of question seeks to delimit or restrict the pupil's answer to a brief and unique phrase, word or symbol. Consequently, an objective-type item can be scored by mechanical devices or by clerks who have no special competence in the subject.

Objective-type items may be further subdivided into two main groupings. In one, the pupil must respond by *supplying* the response. In the second, the pupil responds by *selecting* a response from among those presented to him. The specific objective-type items to be discussed in this bulletin include the completion item, which represents the supply type, and the true-false, multiple-choice and matching items, which represent the selection type.

A. ESSAY QUESTIONS

The Value of Essay Questions

The fundamental contribution of the essay question is that it requires the pupil to develop an answer from his own background and fund of experience, without benefit of suggested possibilities or alternatives, and to express that answer in his own words. Where this is a crucial element in the measurement picture, the use of essay rather than objective questions is indicated. By requiring pupils to present evidence, to evaluate, to analyze and to solve new problems or approach problems in a new way, essay questions can serve to measure some higher level abilities and thus contribute to knowledge about the pupil.

Achieving Reliability of Essay Questions

An important problem in the use of essay questions is achieving reliability of test scores. One source of unreliability in the essay question is subjectivity of rating. Considerable variation of assigned marks has been found, not only between one teacher and another but in ratings by the same teacher from day to day. Usually, extreme cases of subjectivity of rating can be ascribed to failure to use a proper key or to improper rating procedures, but the point remains that the rating of essay answers is inherently a matter of subjective judgment.

A second source of unreliability is possible inadequacy of sampling. When a test includes relatively few essay questions, the sampling of content in the test is likely to be quite narrow. The pupil strong in area A and weak in area B will receive a high score if the test happens

to include an essay in area A and a low score if it happens to include, instead, an essay in area B.

These difficulties should not be interpreted by the teacher to mean that the essay question should be avoided. They do indicate that, when the essay question is used, special care should be taken to achieve reasonable reliability of test scores. In addition, it would seem advisable for the teacher to direct the use of the essay question toward measuring outcomes that objective tests are not believed to measure satisfactorily.

B. COMPLETION ITEMS

Advantages of the Completion Type

In the typical completion item, the pupil is required to complete a statement by writing the answer in the space provided. For example:

The capital of New York State is the city of _____.

The completion item represents a compromise between the recall of the essay item and the objectivity of rating of the true-false and multiple-choice item. Since the pupil is required to write the answer, the emphasis is upon recall rather than recognition. Guessing plays little part. If the items are well drawn, the answers are definite and specific and can be rated quickly and accurately. Wide sampling of content is possible, since a relatively large number of questions can be answered per unit of testing time.

The Problem of Limiting Possible Answers

The chief problem in the construction of completion items is to find questions each of which has a single correct answer. Teachers are often surprised by the variety of correct responses that appear to questions for which they had conceived only one answer. As the number of possible responses to an item increases, scoring becomes more difficult and more time-consuming, and the obtained test scores become less reliable.

Practically speaking, only items involving simple computations or simple verbal associations can be handled satisfactorily in the completion item. The search for such questions too often leads the teacher to concentrate on details. It is rather difficult, in completion items, to measure the higher processes of thought and reasoning. Hence an important limitation of the completion item is that its effective use tends to be restricted to measuring information and specific facts.

C. TRUE-FALSE ITEMS

Advantages of the True-False Type

The true-false item consists of a statement to be judged true or false or a question to be answered *yes* or *no*. For example:

True-False. The capital of New York State is Albany.

This type of item is easily adapted to classroom testing situations. A large number of items can be administered in a relatively brief testing period, so that wide sampling of content is possible. Scoring, of course, is entirely objective.

Limited Applicability of True-False Questions

The true-false item can be used to measure reasoning and understanding but in practice its use is commonly limited to testing information and specific facts. The reason is that true-false items must be based on statements that are absolutely and unambiguously true or false. A relatively small proportion of the significant statements that can be made on any subject satisfy this criterion. Many important outcomes of instruction are generalizations, explanations, evaluations, inferences etc. Since these cannot often be expressed in statements that are universally true, they cannot be tested effectively by true-false items.

The Danger of Exceptions to True-False Statements

To meet the standard of absolute truth, a true-false statement must be so precise in phrasing and so universal in application that it requires no additional qualifications and admits of no possible exceptions. If statements that are only approximately true are presented as items in a true-false test, they pose a difficult and unreasonable problem for the pupil. Not only must he know to what extent the statement is true, but he must also guess what degree of untruth will be tolerated by the teacher.

Usually it is the brighter, better-informed pupil who sees the need for qualifications and the possibility of exceptions to statements presented as true-false items. When this happens, the item loses validity. Consider, for example, the following item, intended as a true statement:

Water boils at 212°.

The better-informed student is likely to mark this item false because it does not specify the Fahrenheit scale, or because it does not specify

standard atmospheric pressure or perhaps because the statement would not be true of sea water. It is this susceptibility of the true-false item to special interpretation by the pupil that, contrary to the popular notion that true-false tests are easy to write, makes the true-false question one of the most difficult to handle effectively.

The Guessing Factor in True-False Questions

Another limitation of the true-false item is its susceptibility to guessing. Without reading the question, the pupil has a 50-50 chance of getting the right answer. This limitation can be overcome, however, by including a large number of true-false items in the test. In a 10-item true-false test, the chances of answering at least 70 percent of the items correctly on the basis of chance alone are about 1 out of 6. If the number of items is increased, the chances are progressively reduced, as follows:

NUMBER OF TRUE-FALSE ITEMS	CHANCES OF ANSWERING AT LEAST 70% CORRECTLY BY CHANCE ALONE
10	1 out of 6
25	1 out of 50
50	1 out of 350
100	1 out of 10,000
200	less than 1 in 1,000,000

It is readily seen that, in longer true-false tests, the guessing factor becomes a minor element in achieving a passing grade.

The Modified True-False Type

A device commonly employed to reduce the guessing element in true-false questions is to have the pupil correct the item if it is false. For example:

> **Directions: If the statement is true, write *true*. If the statement is false, write the word that should be substituted for the italicized word to make the statement true.**
>
> **The capital city of New York State is *Buffalo*.**

This type of item, called the modified true-false, is in reality a cross between the true-false and the completion types. It does not eliminate the guessing element entirely. If a statement is true the pupil still has a 50-50 chance of guessing the correct answer.

D. MULTIPLE-CHOICE ITEMS

Relative Freedom from Any Ambiguity

The multiple-choice item consists of an introductory question or incomplete statement, called the stem, and two or more suggested answers or completions, called the responses. The pupil is instructed to choose the best, or most nearly correct, response. For example:

The capital of New York State is

1. **Syracuse**
2. **Albany**
3. **Rochester**
4. **Buffalo**

Since the multiple-choice form involves selecting the best answer, it avoids the ambiguity associated with application of a standard of absolute truth, which constitutes the serious weakness of the true-false form. The pupil is limited to the choices listed and there is little or no opportunity for him to introduce into the question qualifications or exceptions beyond the intent of the teacher. Thus, suppose that the illustrative true-false item on the boiling point of water, presented above, is rewritten in multiple-choice form:

At what temperature does water boil?

(1) 50° **(3) 182°**
(2) 73° **(4) 212°**

There is no question now as to the correct answer. Failure to specify centigrade or Fahrenheit, while a weakness in the question, is not as serious in this item as in the true-false form. Further, it is readily apparent to the pupil that atmospheric pressure or dissolved solids are not to be considered.

The Guessing Factor in Multiple-Choice Questions

The guessing factor is considerably reduced in multiple-choice questions, especially if 4 or 5 responses are provided. In a 10-item, 4-response multiple-choice test, the probability of obtaining a score of 70 percent on the basis of chance alone is about 1 in 1,000. For a 25-item test, the chance is reduced to about 1 in 1,000,000. To achieve freedom from guessing comparable to that in a 25-item multiple-choice test, a true-false test of 200 items would be required.

Wide Applicability of the Multiple-Choice Form

The broad possibilities of the multiple-choice item are illustrated by a few examples of the types of concepts that can be tested:

1. A _____ is best defined as .....
2. The best reason for _____ is .....
3. What is the cause (or effect) of _____?
4. The essential similarity (or difference) between _____ and _____ is .....
5. Why is _____ important?
6. As _____ changes, what is the change in _____?
7. Under what conditions is _____ true?

It will readily be seen that the multiple-choice form can be adapted to evaluating a great variety of curricular objectives. While it has often been used to measure superficial verbal associations and minor factual details, it can nevertheless be used with great effectiveness to measure reasoning ability and fundamental understanding.

E. MATCHING ITEMS

Advantages of the Matching Type Item

The matching form of objective test exercise consists of a list of premises and a list of responses. The pupil is directed to match one of the responses to each of the premises. For example:

> **In the space next to each opera in Column A write the number preceding the name of the musician listed in Column B who composed the opera.**

Column A	*Column B*
_____ a. *Aïda*	1. Bizet
_____ b. *Marriage of Figaro*	2. Donizetti
_____ c. *Carmen*	3. Gounod
_____ d. *Faust*	4. Mendelssohn
_____ e. *Tristan und Isolde*	5. Mozart
	6. Verdi
	7. Wagner

The matching item requires relatively little reading time on the part of the pupil, as compared with other types of items, and a large number of questions can be asked in a brief period of time. The guessing factor is small. Scoring is simple and objective.

Limited Applicability of the Matching Type Item

While the matching item has some variety of possible applications, in that the lists of premises and responses may consist of names, dates, terms, phrases, statements, references to maps or diagrams etc., it is not as flexible as other item types. Its use is restricted to situations in which the material to be tested lends itself to the listing of a number

of important and related concepts. There are, of course, many significant topics that are unique and cannot be conveniently grouped in homogeneous matching clusters.

III. Writing Effective Test Items

The Art of Writing Test Items

The preparation of test questions is a highly creative and challenging task. Potentially valuable questions, designed to test significant aspects of the curriculum, may lose much of their effectiveness because of flaws in the manner in which they are presented to the pupils. Writing good items is not easy. The pitfalls are many. But if an idea is worth testing, sufficient care should be taken in formulating the test item to assure that it provides a valid and reliable index of what it is designed to measure.

In this chapter a number of suggestions are presented for preparing test questions of different types. It is not possible to discuss all of the possible difficulties and problems of item construction. Consideration is given, rather, to the basic principles that can serve as practical and helpful guides in evaluating and improving local classroom tests.

A. ESSAY QUESTIONS

1. Use the essay question to measure objectives that cannot be measured as well with other question types.

One of the common weaknesses in classroom tests is failure to restrict the essay question to its special purposes, resulting in less efficient testing of objectives that could be better tested in other ways. The essay question should not be used solely to elicit information. Routine testing of information is a function that can be performed more effectively by objective-type questions.

Consider, for example, the following question:

Name and locate 5 important world seaports.

This question does not call for analysis, judgment, problem solving or any of the abilities toward measurement of which the essay question can make a special contribution. It affords little opportunity for written expression. Use of the essay form in this instance serves merely to increase rating time, invite rating problems and make the test scores less meaningful. It may be anticipated that the teacher will have considerable difficulty in deciding whether some of the pupils' borderline answers are really important world seaports. In addition, there may

well be some question as to whether the particular location made by a pupil is sufficiently specific to merit full credit. For example, if a pupil names "Marseilles" as an important seaport, he may give its location as "Europe," "France" or "The Mediterranean." The teacher has no way of knowing whether the pupil is deliberately being vague because he is uncertain of the correct answer or whether he could give a more precise location if he knew what was wanted.

In contrast a question of the following type can be defended as having a special contribution which justified the essay approach:

> **"Today we are better prepared to combat an economic depression than we were in 1929." Do you agree or disagree with this statement? Why?**

From his background of knowledge and experience the pupil must select the pertinent information, evaluate it, decide how best to express it and write the answer. The mental processes the pupil goes through would appear to be basically different from those that he would use in answering related questions presented in objective form.

2. Limit and define the freedom of pupil responses to essay questions.

Freedom of response is a fundamental characteristic of the essay question, but the teacher should define the limits within which this freedom may operate. Too often an essay question requires, in effect, that the pupil write all he knows about a broad, general topic. An example of such a question, taken from a teacher examination, is the following:

> **Describe the development of transportation and communication from the period of prehistoric man to the present time.**

The difficulty with this type of question is that it is impossible to compare validly the widely different discussions one is likely to obtain in answer to such a question. The fundamental assumption in any test is that it constitutes a uniform work sample on which pupils may be compared. Questions of the above type are only deceptively uniform. Actually, there are as many different questions as there are pupils, depending upon their subjective interpretations of the question. Hence inferences concerning the comparative achievement or progress of the pupils are grossly unwarranted.

The problem of delimiting the scope of pupils' responses is not solved merely by reducing the length of the answer required. Brief essays may also allow excessive freedom of response. For example:

> **Write a paragraph on the topic *Communism.***

Pupils may choose to discuss a large variety of different topics relating to communism.

To achieve comparability of essay scores, the pupils must be directed into a specific situation and allowed to react freely within the limitation thus imposed. For example:

> **A person will feel more uncomfortable on a hot, humid day than on a hot, dry day. What is the biological explanation?**

This question allows the special advantages of the essay question to come into play. It provides an opportunity to analyze a problem, formulate a logical answer and express it in correct English. Yet it is relatively narrow in scope. All pupils will be writing on the same topic and their answers will be more directly comparable.

If the longer essay is used, it should be framed around a specific problem, delimited in scope and subdivided into major points for consideration. For example:

> **Show how each of the following has been a factor in the development of democracy in the United States:**
> **(1) the characteristics of life on the western frontier**
> **(2) the public schools**
> **(3) newspapers and periodicals**

This item provides a frame of reference for the pupils. They are not invited to roam aimlessly through an uncharted wilderness. A minimum degree of comparability of answers is insured by the fact that all pupils are directed to consider the same major elements in making their analysis.

A helpful procedure for the teacher to follow is to outline beforehand all the possible answers that pupils may write in answer to an essay question. Sometimes it becomes immediately apparent that there are too many highways and byways down which the pupils may travel, indicating the need for delimiting the scope of the question. Sometimes the reverse is found to be the case—that there is really not very much the pupil can give in answer to the question as framed. In such instances the scope of the question can be broadened appropriately.

3. Use many brief essays rather than one or two extended essays.

There are two important reasons for using brief essays in preference to long essays. First, a test composed of many brief essays will permit a wider sampling of the content of the course and so will provide a more accurate picture of the pupil's total achievement. Second, the test score will be more reliable. If the essay test contains only a few questions, each calling for a long answer, the pupil who happens to mis-

interpret one question will be penalized disproportionately to his real knowledge of the subject. To be reliable, a test must provide for a reasonable number of independent responses. A long essay may be carefully delimited in scope by subdivision into component elements, but these elements are generally too closely related to provide either the breadth of sampling or the higher reliability that can be obtained with an equivalent number of independent brief essays.

4. Indicate clearly in each question the desired extent and depth of answer.

Unless the purpose of an essay question is to determine the depth or scope of analysis that the pupil will make without specific direction, the desired scope of the analysis should be indicated by the question. Consider the following example:

> **Explain how the Federal Government has aided the development of business in the United States.**

Sometimes a teacher will feel that a question of this type is answered adequately only if the pupil gives a good discussion of two different instances of Federal aid to business. If the pupils are not made aware of this fact, the results may be unfortunate. A pupil may discuss only one instance of Federal aid and decide to stop there, although he might well have given two, three or even four if requested to do so. On the other hand, another pupil might go on at great length and discuss seven or eight different points, far more than necessary to achieve a perfect mark in this question, thus wasting valuable time that might have been spent more profitably in answering other questions. If the scoring key will allow full credit only for discussing two instances, the student should be so informed:

> **Explain *two* ways in which the Federal Government has aided the development of business in the United States.**

Similar considerations should govern the selection of the keywords commonly used in essay questions: "explain," "describe," "outline," "list," "state" etc. The teacher should decide beforehand exactly what level and type of treatment is desired for a perfect answer and then select the appropriate terminology that will lead the pupils to provide such treatment in their answers. Consider, for example, the following question:

> **State three basic problems that the President of the United States is likely to face next year.**

For the teacher to penalize the pupil who states the problems baldly and without any explanation would be manifestly unjust, for the pupil

has done all the question asks of him. If the teacher wishes to base the pupil's score on the extent to which the pupil realizes the implications of these problems, or their possible solutions, or their importance to the nation, the question should be revised to call for such additional depth of analysis.

B. COMPLETION ITEMS

1. Avoid the loose, ambiguous item that does not tie down the answer to one or two specific words or phrases.

Every completion item should be scrutinized carefully to make certain that it calls for a specific, limited answer. Too often the completion item offers ample opportunity for pupils to misinterpret the question or to evade it. For example:

The population of the United States is _____.

The teacher should not be at all surprised to find a large variety of answers—"increasing," "largely urban," "highly literate," "prosperous," "enumerated every 10 years," "growing older" or "about 50% men." Drawing the line between what does and what does not constitute an acceptable answer to an ambiguous question can be a knotty problem.

Another example of a loosely framed item is:

S. T. Coleridge wrote _____.

This question is so framed that any work of Coleridge would be a correct answer. Reference by the pupil to an unusual title might well necessitate extensive research by the teacher. If the purpose of the item is to determine whether the pupil knows that Coleridge wrote "The Ancient Mariner," the question should be directed at this point. In many cases use of the direct question will help further to tie down the answer:

Who wrote "The Ancient Mariner"? _____

2. Do not require more than one or two completions to be made in any one item.

Sometimes completion items are framed so that two or more blanks are left in the same sentence or paragraph. As the number of blanks increases, the interrelationships and combinations become more complex. Such questions frequently require an excessive amount of time to answer, require careful scoring for part-credit allowance and introduce the extraneous factor of verbal facility into the measurement pic-

ture. There are intelligence tests built entirely on the principle of completing sentences In the following item, for example, too many words have been removed to permit the items to pose a clear-cut problem:

> **Most (*green*) plants produce (*sugar*) from (*water*) and (*carbon*) dioxide.**

3. In general, place the blank near or at the end of the statement.

If the blank to be filled in by the pupil comes at the beginning of the statement, the pupil must begin with a blank, read the statement, and then retrace his steps to decide what should be written in the blank. For example:

> **_____ are the hairlike structures by means of which paramecia move.**

The item can be grasped by the pupil more readily if it entails a direct and simple approach. By the time the pupil reaches the blank space he should have all the information he needs to write the answer without further ado—if he knows it.

> **Paramecia move by means of hairlike structures called _____**

4. Avoid extraneous clues to the correct answer.

Sometimes the grammatical structure of a completion item provides a clue to the correct answer. For example:

> **A fraction whose numerator is greater than its denominator is called an _____ fraction.**

The use of "an" in this item would indicate to an alert pupil that "proper" cannot be the correct answer, for "an proper fraction" would be incorrect English. While the teacher may not want to discourage alertness in the pupil, the purpose of the test is, after all, to measure mathematical ability. This item is easily improved by revising the item to eliminate the article:

> **Fractions whose numerators are greater than their denominators are called _____ fractions.**

Another type of extraneous clue that is not uncommon is the use of short blanks for short words and long blanks for long words. The length of the blank should not indicate to the pupil the length of the word or phrase to be inserted. Nor should the number of blanks indicate the number of words omitted. In the following item, for example,

a pupil is not likely to answer "oxygen" or "nitrogen" in the face of indications that the correct answer has a compound name:

The gas that makes a cake rise is _____ _____.

5. In general, do not use the completion form when no recall is involved.

The basic principle underlying the completion item is that, since the pupil is required to write the answer to a question, the emphasis is upon recall rather than recognition. Sometimes, however, a question is a completion item in form only. For example:

The density of a floating body is _____ than 1.

Essentially this item is no different from a true-false item or a two-option multiple-choice item. There are only two possible answers. In such instances the completion form offers no special advantage. No recall is involved. The guessing element is not reduced. Items of this type, especially if there are a number of them in the test, might well be converted into more objective types for more efficient use of testing time and greater objectivity in scoring.

6. In computation problems, specify the degree of precision expected.

When the emphasis of a mathematics problem is upon understanding and method, the problem may frequently be arranged to come out evenly. When the answers involve fractions, decimals or approximations, however, the pupil should be advised of the standard of precision on the basis of which he is to be marked. He cannot be expected to answer an item correctly if he does not know what is expected of him. Consider the following item:

If the radius of a circle is 1 inch, the area is _____.

It would be unfair to mark the answer "3" or "3⅛" wrong on the grounds that the answer should be 3.14. Similarly, it would be unfair to mark the answer 3.14 wrong because the answer expected is 3.14 *square inches*. If the correctness of a numerical response depends upon stating the unit of measurement, this fact should be made clear. If not, it is best to include the unit of measurement in the statement of the question.

The radius of a circle is 1 inch. Its area, correct to two decimal places, is _____ square inches.

C. TRUE-FALSE ITEMS

1. Avoid loosely worded and ambiguous statements.

Sometimes a true-false item can be interpreted in different ways. The pupil who recognizes the ambiguity is faced with the task of guessing which meaning is intended. If he guesses wrong, he may lose credit although he knows the answer to the question intended by the teacher. For example:

The earth is nearest the sun in December.

The intended answer is *true*, on the basis that December is the approximate time of year when the earth, moving in its orbit, approaches the sun most closely. The way the item is phrased, however, it may be interpreted to mean that in December the planet nearest the sun is the earth. This is false. The teacher should constantly be alert to such possibility of ambiguity in even the simplest of true-false statements.

Another example of an ambiguous true-false statement is the following:

All men are created equal.

What is the intended answer in this case? The Declaration of Independence affirms that all men are created equal and we consider all men equal before the law. Is that the point of the question? Or is the point, rather, that men are not equal in such factors as mental and physical capacity? The pupil must try to guess what the teacher has in mind. The consequences of ambiguity are to confuse the pupil and to impair the value of the test.

2. Base true-false items upon statements that are absolutely true or false, without qualifications or exceptions.

The problem of exceptions to true-false statements is related to ambiguity because, again, the pupil is uncertain as to what the teacher wants. In this instance, however, confusion arises in the mind of the pupil not with relation to the central point of the question but with relation to the depth of analysis required and the strictness of standards of truth that must be applied. The following statement is generally true and will be accepted by most students at the superficial level at which it is intended:

Animals that fly are called birds.

Some superior students, however, are likely to think that this is a more profound question designed to test whether they realize that

some birds do not fly and that some flying animals are not birds. On the basis of these exceptions, they are likely to mark the item "false."

The implication for the teacher is that every true-false statement should be reviewed critically in an attempt to anticipate possible exceptions or qualifications that may reasonably arise in the minds of the pupils.

3. Highlight the central point of the question by placing it in a prominent position in the statement.

A true-false item should be so framed that the central point of the question will receive adequate attention on the part of the students. It is not wise to try deliberately to slip the keywords of the item past the pupil. To do so tends to defeat the purpose of the question. A pupil may get the item wrong because of careless reading, although he may possess the knowledge or understanding that the item is designed to measure.

In the following item, the critical point is subordinated in a parenthetical clause:

> **The 16th amendment to the Constitution, ratified in 1920, authorized Congress to collect taxes on incomes.**

The intended answer is "false" because the 16th amendment was ratified in 1913. Assuming that the major point the teacher wants to measure is the ratification date, the item can easily be improved by reversing the clauses, thus making the crucial element more prominent:

> **The 16th amendment to the Constitution, which authorized Congress to collect taxes on incomes, was ratified in 1920.**

On the other hand, if the teacher is more interested in the substance of the amendment than in the ratification date, then the date, if given at all, should be given correctly:

> **The 16th amendment to the Constitution, ratified in 1913, authorized Congress to collect taxes on incomes.**

4. Avoid double-barreled and multi-barreled statements that are partly true and partly false.

Sometimes the important point being measured is obscured not because of its subordinate position in the statement but because two, three or four different ideas are combined in a single item. For example:

> **Shakespeare wrote *Hamlet*, the story of a prince of Denmark, and he died in the 18th century.**

A common belief is that incorporating a number of different ideas into a single item serves to make the test more comprehensive. This is not necessarily the case. The pupil may still answer the item on the basis of familiarity with only one of the ideas included. Multiplicity of ideas serves rather to confuse the measurement picture. In the above item, for example, pupils may well believe they have grasped the central point of the question after they have read the first part, tending to overlook entirely that the date of Shakespeare's death is also an important point for them to consider. Better range of sampling and more valid measurement would be obtained by splitting this one item into three separate parts, each testing a single major idea.

5. Avoid the use of long and involved statements with many qualifying clauses.

The difficulty with long, involved statements is, again, that the pupil is unable to identify readily the crucial element in the item. For example:

> **Pauperism is a relative term but as generally employed it means that economic and social state in which persons have not sufficient income to maintain health or physical efficiency.**

The classroom test should seek to provide an achievement score, relatively uncontaminated by such factors as reading ability and verbal intelligence. Both of these extraneous factors would seem to be heavily involved in grasping the central point of a complex item such as the one cited above. Whether the pupil understands what is meant by pauperism is probably better measured by some more simple and direct statement, such as the following:

> **A pauper is a person with insufficient income to maintain health or physical efficiency.**

6. Avoid trick questions.

An example of the trick question is the following:

> **The chief component in the automobile engine combustion chamber is nitrogen.**

The answer is intended to be "true" (despite the fact that nitrogen is an inert gas that does not enter into combustion) because nitrogen constitutes 80 percent of the air drawn into the combustion chamber. Most students will, of course, think only in terms of gasoline and oxygen and will get the item wrong.

The trouble with the trick question is that, generally, it measures something quite different from the objective of the test. What this something is, we can never be quite sure. It may be alertness, or intelligence, or perhaps the tendency to be suspicious. In any case, the net result of items of this type, in addition to the harmful effect they are likely to have on pupil attitudes toward examinations and on pupil-teacher relationships, is to impair the value of the test as an achievement measure.

7. Avoid "window dressing" that may serve to confuse the pupil.

"Window dressing" is a term applied to information in a test item that is not essential to answering the item but is designed to make the item more interesting to the pupil, to make it more realistic or practical or to serve an instructional purpose. These objectives are commendable. Caution must be exercised, however, that the window dressing does not get in the way of the measurement objective. Excessive window dressing may result in a test that, as a whole, places too much emphasis upon reading speed and comprehension.

In individual items, the introduction of extraneous material may impair item validity by obscuring the central point of the item:

> **Since we live in an age where everything is measured, that is, we buy meat by the pound and we drive so many miles, we must also have a unit with which to measure angles; this unit is called a gram.**

Many pupils will have difficulty separating the window dressing from the point being tested. The item could be presented much more simply and effectively as follows:

> **Angles are measured in units called grams.**

On occasion the window dressing itself is the direct source of ambiguity and confusion. For example:

> **Every schoolboy who reads the "funnies" knows that radar deals with the detection of unknown distant objects.**

In this instance the effect of the window dressing is to ruin the item. No pupil can validly be marked wrong for denying that every boy who reads the "funnies" knows about radar.

8. Avoid negative questions whenever possible, and double negatives in any case.

There are instances in which negative questions are desirable and effective. If the concept to be tested is essentially a negative one, for example, it can probably be tested as well in no other way. Thus:

> **Iodine should *not* be applied around the eyes.**

In such cases it is desirable to emphasize the negative nature of the item by underlining the negative words or by using distinctive type styles.

Indiscriminate use of negative questions, however, has serious limitations. Negative questions often complicate meaning needlessly, tend to create confusion by requiring a radical shift in orientation from previous questions and reduce reliability by inducing careless errors due to overlooking the negative. To ask a pupil to conclude that a negative item is false because it is not true that something is not true involves a twist in reasoning that can be extremely confusing.

Usually the instructional objective of the teacher can be approached just as well, and measured much more effectively, by the more direct approach. Consider, for example, the following item:

> A United States Senator is not elected for a 6-year term.

This same concept can be tested more directly in a positive manner, whether the correct answer is to be "true" or "false":

> A United States Senator is elected for a 6-year term.
> A United States Senator is elected for a 2-year term.

Items containing double negatives are especially likely to be troublesome:

> A bill vetoed by the President does not become law if not subsequently approved by a ⅔ vote of each house of Congress.

Again, the same idea can easily be tested in a positive form:

> A bill vetoed by the President will become a law if subsequently approved by a ⅔ vote of each house of Congress.

9. **Avoid extraneous clues in the form of specific determiners.**

Analysis of the usual true-false examination will reveal that the pupil who arbitrarily goes through the examination and places a "false" next to every item containing an extreme word (such as *always, never* and *all*) and a "true" next to every statement containing a qualifying word (such as *usually, sometimes* and *may*) will be right more often than he will be wrong. These clue words, or specific determiners, are used by teachers so frequently because they are found necessary to make the items foolproof—either absolutely true or absolutely false. Pupils soon become alerted to these keywords and base their answer on the specific determiner rather than on the item as a whole. Examples of items with such extraneous clues are the following:

Cuticles should *always* be cut.
Malnutrition *never* occurs in families that have ample funds to spend for food.
Clothes *usually* dry more quickly on a windy day than on a calm day.

The point is not that qualifying words must never be used in true-false items. It may not be possible or desirable to avoid them entirely. What should be avoided is the pattern of having certain types of words usually go with certain answers. If items containing *always* or *usually* are sometimes true and sometimes false, the pupils are not provided extraneous clues to the correct answer.

10. In the modified true-false type of item, indicate the word or words that should be corrected.

In the modified true-false item type, the pupil is required to correct a statement if it is false. A common practice is to give no clue to the pupil as to what words should be changed to correct the item.

There are a number of objections to this practice. First, it invites scoring difficulties. There may be several methods of improving the item, so that several different answers may be possible. For example:

The famous Parthenon still amazes visitors to Rome.

Some pupils, instead of substituting Athens for Rome, will substitute the Colosseum, or some other building, for the Parthenon. This may require not only considerable research on the part of the teacher, but also some rather fine decisions as to what may or may not be considered famous. Difficulties of this type are easily avoided by underlining Rome and requiring the pupils, if the statement is false, to substitute for that word.

Second, not all pupils will be answering the same item or taking the same test. The question "Where is the Parthenon located?" is quite different from the question "Name a famous building in Rome." Unless pupils take the same test, their scores are not directly comparable.

Third, the time required to answer the items is increased. For example:

A straight line touching a circle at but one point is a segment.

The pupil must pause in turn at each of the significant words in the statement and weigh the effect of a possible change in that word. This may be a laborious and time-consuming procedure. If "segment" were underlined, the central point of the question could be grasped much more quickly. In the time saved perhaps two or three additional questions could be answered.

Finally, the very task of analyzing the statement and deciding the proper point at which to make a change probably involves a strong element of verbal facility and general intelligence, which is extraneous to the fundamental purpose of the test and so operates to lower its validity.

D. MULTIPLE-CHOICE ITEMS

1. Use either a direct question or an incomplete statement as the item stem, whichever seems more appropriate to effective presentation of the item.

Some item ideas can be expressed more simply and clearly in the incomplete statement style of question. On the other hand, some items seem to require direct question stems for most effective expression. Thus, neither of the following items can be presented quite as neatly in the other forms.

Which is a mammal?

1. **rat**
2. **mosquito**
3. **rattlesnake**
4. **spider**

Flowers are necessary to produce

1. **bulbs**
2. **roots**
3. **seeds**
4. **cuttings**

Teachers should use the item style that seems most appropriate for the particular item in hand. If, in a given instance, it seems to make little or no difference which item type is used, the teacher should choose the style that she can generally handle most effectively. There is no evidence that either type is inherently superior to the other.

Teachers who have not had experience in writing multiple-choice items may find that, at the beginning, they will tend to produce fewer technically weak items when they try to use direct questions than when they use the incomplete-statement approach. The reason is that in the incomplete-statement approach it is often difficult to arrange qualifying phrases or words to produce a perfectly clear statement. In addition, because of its specificity, the direct question induces the item writer to produce more specific responses.

2. Items should be written in clear and simple language, with vocabulary kept as simple as possible.

Like any other item, the multiple-choice item should be perfectly clear. Difficult and technical vocabulary should be avoided unless essen-

tial for the purpose of the question. Sentence structure should be as simple as possible. Complex sentences should be broken up into two or more separate sentences. Qualifying phrases should be placed near the terms they qualify. The important elements should generally appear early in the statement of the item, with qualifications and explanations following.

In some instances the meaning of an item is not too easy for the pupil to grasp, and occasionally it may be well obscured. The following item would be considerably less confusing to pupils if the stem were re-written along the lines indicated:

(Poor) The great circle, one-half of which is the Greenwich Meridian, is on the other side of the

(1) International Date Line
(2) Equator
(3) Tropic of Cancer
(4) Tropic of Capricorn

(Improved) The Greenwich Meridian is one-half of a great circle. What do we call the other half?

Another example of an item that is unnecessarily confusing because of language difficulties is the following:

(Poor) In New York State, the season of the year in which records indicate maximum statistical occurrence of hail-storms is

1. fall
2. late winter
3. midwinter
4. summer
5. very early spring

(Improved) In New York State, hailstorms are most likely to occur in .....

The power of an item to discriminate between the competent and incompetent may be seriously limited by lack of clarity. Except in the case of certain types of intelligence or reading test items, the difficulty of an item should arise from the problem involved rather than from the language in which it is expressed. Test items should not be verbal puzzles. They should indicate whether the pupil can produce the answer, not whether he can understand the question.

3. Each item should have one and only one correct answer.

While this requirement is obvious, it is not always fulfilled. Sometimes teachers produce items involving issues so controversial and debatable that even competent authorities are unable to agree on one correct answer. More often the trouble is failure to consider the full implications of each of the responses. In the illustration below, for

example, "writer" is a possible correct answer not intended by the question author. The item is improved by "spoiling" this response so that it is definitely incorrect.

> **(Poor) H. L. Mencken has achieved his greatest fame as a (1) writer (2) biographer of Abraham Lincoln (3) poet (4) psychologist (5) student of the American language**
>
> **(Improved) H. L. Mencken has achieved his greatest fame as a (1) writer of popular novels . . . . . .**

The multiple-choice item allows greater control and direction of pupil response than does the true-false item, but there is still ample opportunity for unforeseen pupil interpretation leading to more than one correct response. For example:

> **Choose the man who does not belong in this group.**
> 1. **Leonardo da Vinci**
> 2. **Raphael**
> 3. **Dante**
> 4. **Rembrandt**

The intended classification is probably "painters," making Dante the answer. However, another possible classification is "Italians," in which case Rembrandt is the answer. In any classification item of this type, it is always important to consider the possibility of some unforeseen classification presenting itself to the pupils.

4. Base each item on a single central problem.

The multiple-choice item functions most effectively when the pupil is required to compare directly the relative merits of a number of specific responses to a definite problem. Sometimes an item consists merely of a series of unrelated true-false statements, all of which happen to begin with the same phrase:

> **A large deficit in the national budget**
> 1. **is likely to result in inflation**
> 2. **occurs usually in times of prosperity**
> 3. **can be avoided by establishing the gold standard**
> 4. **indicates an unfavorable balance of trade**

In reality this item is concerned with four different problems. As a series of true-false statements the question loses the central advantage of the multiple-choice form. The pupil is not really required to choose the best of a number of alternative choices to a single problem, but rather to decide which of a number of independent true-false statements is more true than the others. This is a difficult and confusing task at best. Subjectivity of interpretation and possible ambiguity of each of

the true-false statements is as great a danger here as in the standard true-false item.

The item is improved by delimiting the scope of the question to the consequences of a large deficit. The pupil compares directly the relative merits of a group of homogeneous alternative responses to this specific problem. He does not have to determine the relative degree of absolute truth of four independent statements.

A large deficit in the national budget is most likely to result in

1. **inflation**
2. **high interest rates**
3. **low profits**
4. **an unfavorable balance of trade**

Sometimes an item does not consist of entirely unrelated true-false statements but still involves more than one central problem. For example:

In case of sunstroke, the victim

1. **should have his head lowered**
2. **has a weak pulse**
3. **should be given a stimulant**
4. **should have cold water poured over him**
5. **has a pale face**

This item involves two separate problems: the symptoms of sunstroke and treatment for sunstroke. Items of this type are subject to the same general criticism as a collection of unrelated true-false items. The pupil is looking for a single problem and he becomes confused and uncertain about the main point of the question. In addition, the merits of the "best" solutions to entirely different problems are not always directly comparable.

The item is improved by focusing the problem on either symptoms or treatment. An incidental advantage is that the teacher now has two items for possible use instead of one.

Which is a symptom of sunstroke?

1. **weak pulse**
2. **pale face**
3. **high temperature**
4. **excessive perspiration**
5. **severe chills**

A proper first aid treatment for a victim of sunstroke is to

1. **lower his head and shoulders**
2. **apply a cold compress to his head**
3. **cover him with a blanket**
4. **administer a warm liquid stimulant**
5. **administer artificial respiration**

5. State the central problem of the item clearly and completely in the stem.

The stem or introductory statement of an item should be meaningful in itself and should convey clearly the central problem of the item. It should not be necessary for the pupil to read and reread all the responses before he can understand the basis upon which he is to make a choice. The following item, for example, is concerned with a single problem, but there is nothing in the stem to indicate to the pupil what that problem is. Just what does the teacher want to know about *The Yearling*? That can be inferred only by reading all the responses.

The Yearling
1. is a story of the Ozark foothills
2. describes life in Washington timber country
3. has as its setting the Allegheny Mountains
4. has the Florida scrub as its locale

The item can be improved by stating the problem in the stem, so that the pupil knows precisely what to look for before he reads any of the responses. In a test made up of such items, the pupil can proceed more rapidly and with greater confidence.

The locale of *The Yearling* is the
1. Ozark foothills
2. Washington timber country
3. Allegheny Mountains
4. Florida scrub

6. In general, include in the stem any words that must otherwise be repeated in each response.

The stem should contain everything the answers have in common or as much as possible of their common content. This practice serves to make the item shorter, so that it can be read and grasped more quickly. Transferring common elements from the responses to the stem also serves to highlight the differences among the responses and so to focus attention on the central issue. In the following illustration, the alternatives contain material that may well be incorporated into the stem.

(Poor) Milk can be pasteurized at home by
(1) heating it to a temperature of 130°
(2) heating it to a temperature of 145°
(3) .....

(Improved) Milk can be pasteurized at home by heating it to a temperature of (1) 130° (2) 145° (3) .....

It is not always possible, or desirable, to eliminate all words common to the responses. If the retention of common words in all of the re-

sponses would serve to make the item easier to understand, they should be retained. For example:

The arteries and the capillaries are part of the
1. circulatory system
2. excretory system
3. digestive system
4. nervous system

7. Avoid negative statements.

Negative statements in multiple-choice items lead to the same difficulties and confusions as in true-false items. In the case of multiple-choice items, special care must be exercised against the double negative. In the mind-set of the question writer, a combination of negatives in the stem and in the responses may not be especially troublesome. The pupil, however, is likely to experience considerable difficulty, especially in the case of items involving complex or abstract ideas:

Which statement concerning antibodies is not true?
(1) The same organism cannot cause the production of two kinds of antibodies.
(2) Antibodies will not destroy white blood corpuscles.
(3) .....

The item is improved by employing a positive approach:

The effect of an antibody in the blood is to destroy the
(1) organisms causing the production of that antibody
(2) white blood corpuscles
(3) .....

8. Avoid excessive "window dressing."

As in the case of other item types, the item should contain only material relevant to its solution, unless selection of what is relevant is part of the problem. In multiple-choice questions there is less possibility than in true-false items that the irrelevant material may introduce ambiguous elements that will lead the pupil to give the wrong answer. However, pupils vary widely in speed of reading and in reading comprehension. The inevitable advantage given to the good reader in most objective tests should not be increased unnecessarily.

(Poor) Sue helps her mother plan meals for the baby. They choose only foods that will be healthful and will help avoid illness. Which illness is prevented by including orange juice in the baby's diet?
(1) scurvy (2) pellagra (3) rickets (4) diabetes
(Improved) Orange juice is included in the diet of infants primarily to help prevent .....

9. Place the choices at the end of the incomplete statement.

A practice sometimes followed in teacher examinations is to insert the choices or responses in the middle of a statement. For example:

> **A (1) transformer (2) condenser (3) motor (4) choke coil will step up the voltage of a current.**

This practice tends to be confusing. At the time he reads the choices the pupil has no clear idea of what the problem is. Sometimes it is difficult to separate the choices from the stem, especially if the choices are long or complex. Placing the choices at the end of the stem makes for continuity of reading and better understanding.

> **The voltage of a current can be stepped up by means of a**
> 1. **transformer**
> 2. **condenser**
> 3. **motor**
> 4. **choke coil**

10. Make the responses grammatically consistent with the stem and parallel with one another in form.

Unless the responses are grammatically consistent with the stem, the pupil will have difficulty in grasping the meaning of the question. In the following item, none of the alternatives is grammatically consistent with the stem in the sense that it completes a clearly expressed and logical idea.

> **Most pedestrian deaths in traffic accidents in New York State occur**
> 1. **it is an intersection protected by a stop sign**
> 2. **there is a traffic light**
> 3. **there is no traffic signal at all**
> 4. **crossing between intersections**

It is absurd, for example, to say that pedestrian deaths occur crossing between intersections. Lack of grammatical consistency in this item is an indication of cloudy thinking and a source of considerable confusion. The item is made more meaningful by making the alternatives consistent with the stem:

> **Most of the pedestrians killed in traffic accidents last year in New York were crossing the street**
> 1. **at intersections without any traffic signals or controls**
> 2. **at intersections protected by stop signs**
> 3. **at intersections protected by traffic lights**
> 4. **in the middle of the block between intersections**

In addition, expressing the alternatives in parallel form brings the central point of the question—where accidents occur—more sharply

into focus. The pupil is helped to compare the distinguishing elements of the different alternatives. As a result, he can arrive at his answer more rapidly and with less opportunity for careless error.

11. Make all responses plausible and attractive to pupils who lack the information or ability tested by the item.

The incorrect responses should be plausible answers. So far as possible, each response should be designed specifically to attract pupils who have certain misconceptions or who tend to make certain common errors. In the following illustration, responses 4 and 5 are implausible. The item is improved by substituting responses more closely related to the others in meaning.

(Poor) What are electrons?
(1) neutral particles
(2) the nuclei of atoms
(3) negative particles
(4) types of buildings
(5) mechanical tools

(Improved) What are electrons?
(1) neutral particles
(2) the nuclei of atoms
(3) negative particles
(4) neutralized protons
(5) radiated particles

12. Arrange the responses in logical order, if one exists.

Where the responses consist of numbers, they should ordinarily be arranged in ascending order. This practice serves to make it considerably easier for the pupil to locate his choice and mark his answer accurately.

An exception is often made when the responses include some of the numbers 1 to 5. There will be a strong tendency for the pupil to confuse the absolute value of the answer with the response position used to indicate it. Thus, in the following item, the pupil may well give the response "4" instead of "2" because the numerical answer is "4."

The square root of 16 is
(1) 1 (2) 4 (3) 8 (4) 16

Even if the choices were lettered the pupil might think of them numerically and give the answer D instead of B because D is the fourth letter. Hence it is generally considered better, in such cases, to have the numerical responses coincide with the response positions used to indicate them. For example:

The square root of 16 is
(1) 1 (2) 8 (3) 16 (4) 4

If an item contains one or more pairs of responses dealing with the same concept, they should usually be placed together. In the following item, it is preferable to arrange the responses as shown rather than to distribute them at random among the choice positions:

Consumer prices in 1954, as compared with 1939, have

1. **decreased about 50%**
2. **decreased about 10%**
3. **increased about 10%**
4. **increased about 50%**
5. **increased about 100%**

13. Make the responses independent and mutually exclusive.

Responses should not be interrelated in meaning. Sometimes two or three of the responses may cover the entire range of possibilities, so that one of them must necessarily be correct. Sometimes one response may include one or more of the other responses, so that they must all necessarily be false. Interdependence of responses helps the student to eliminate wrong answers and reduces reliability by decreasing the number of effective, functioning responses. In the illustration, the first three responses cover the entire range of possibilities. The item is improved by making the responses independent.

(Poor) As a balloon rises in the air, the (1) atmospheric pressure *increases* (2) atmospheric pressure *decreases* (3) atmospheric pressure remains *the same* (4) volume of the balloon decreases (5) weight of the balloon increases

(Improved) The volume of a balloon increases as the balloon rises in the air because the (1) atmospheric pressure decreases (2) weight of the balloon increases (3) density of the air in the balloon increases (4) temperature of the air in the balloon decreases (5) effect of gravitation on the balloon decreases

Another common example of interdependent responses is the following:

Approximately what percentage of the population of New York State lives in New York City?

(1) less than 20% **(3) more than 40%**
(2) less than 40% **(4) more than 60%**

The item is, in effect, a two-response item. If 1 is correct, 2 is also correct; if 4 is correct, 3 is correct. Responses 2 and 3 between them cover the entire range of possibilities. Hence the alert pupil can immediately eliminate 1 and 4 from consideration. The item can be improved in a number of ways that would serve to make the re-

sponses completely independent of each other. The simplest approach is probably the following:

Approximately what percentage of the population of New York State lives in New York City?

1. less than 20% **3. between 40% and 60%**
2. between 20% and 40% **4. more than 60%**

14. Avoid extraneous clues.

Since in the multiple-choice item the pupil is required to associate one of several alternative responses with the stem, any aspect of the question that provides an extraneous basis for correctly associating the right answer or for eliminating a wrong response constitutes an undesirable clue. Such clues may assume a variety of forms. They may be based upon grammatical consistency or inconsistency between alternatives and stem, upon repetition of keywords in the stem and the correct answer, upon rote verbal association between the stem and the correct answer, upon the unusual length or construction of the correct answer etc. It requires considerable experience and skill upon the part of the teacher to avoid such extraneous clues in items.

Consider the following item:

Who played the most important part in establishing the doctrine that European nations should not interfere in the Western Hemisphere?

(1) Monroe (2) Hamilton (3) Jackson (4) Franklin

The word "doctrine" in the stem points immediately on the basis of rote association to the correct answer, "Monroe," without any reference to the substance of the doctrine. This irrelevant clue can be eliminated by rewording the stem so that the correct association must be made on the basis of the whole idea rather than a part:

Who played the most important part in establishing the principle in American foreign policy that European nations should not interfere in the Western Hemisphere?

In the following item, only one of the alternatives looks like a French name:

The Frenchman who first explored the St. Lawrence River was

(1) Jacques Cartier **(3) John Cabot**
(2) Sir Walter Raleigh **(4) Henry Hudson**

The item can be improved by rewording the stem as follows:

The St. Lawrence River was first explored by .....

Frequently, as the result of an attempt to make the correct answer complete and comprehensive, this response is markedly longer than the others. The following is an example.

> **Candy should be eaten sparingly between meals because it (1) causes diabetes (2) supplies an excess of energy (3) causes worms (4) stimulates bile (5) dulls the appetite for foods rich in other necessary elements**

Alternative 5 is much longer than the others. It can be improved simply by eliminating the additional bolstering information.

> **(5) dulls the appetite**

It is well to avoid in a test any consistent tendency to have the correct answer, as compared with the other responses, generally longer or shorter, more specific or more abstract, more technical or less technical etc. Short of a pure guess, correct association between stem and correct answer should be possible only on the basis of relevant knowledge or understanding.

15. Use the "none-of-these" option with caution.

The "none-of-these" option is used in what is sometimes called the "open-end" item in an attempt to retain some of the merits of recall type items. For example:

> **The leader of the gods in Greek mythology is**
> 1. **Apollo**
> 2. **Ares**
> 3. **Pluto**
> 4. **Zeus**
> 5. **none of these**

Actually, it is questionable whether the "none-of-these" option accomplishes this purpose very effectively. Answering the item may well remain entirely a process of recognition. The pupil scans the listed answers. If he recognizes the answer, he chooses it. If he does not, he chooses "none of these." Further, the guessing factor remains unchanged. In a five-option item the pupil still has a 1 in 5 chance of guessing the correct answer.

There may be particular circumstances where the "none-of-these" response would have special value. "None of these" may be a useful response in an item designed to measure the pupil's ability to recognize a situation in which no definite conclusion is to be drawn, such as the following:

In an experiment, two mice are fed the same food except that one is given milk and the other is given coffee. The mouse that is given coffee dies. What does this experiment prove?

1. **Coffee is poison to mice.**
2. **Everyone should drink milk.**
3. **Milk is a better food than coffee.**
4. **None of these is proven.**

Sometimes "none of these" is used in mathematics problems where there is some fear that listing the answers will enable the pupil to work the problem backward.

Which number is the largest prime factor of 418?
(1) 11 (2) 2 (3) 3 (4) 29 (5) none of these

If the item listed five alternative numbers, the pupil could determine the correct answer by trying each alternative in turn, without having to find the prime factors of 418. In the present form of the item, trial and error provides no clue to the correct answer—19. (Of course, when "none of these" is used in a test it should be used a number of times, and not always as the correct answer, or its use will constitute an extraneous clue.)

To counteract these special advantages, the "none-of-these" option has a number of serious limitations. One danger is that a pupil may correctly choose "none of these" but choose it for the wrong reason. In the illustration above, for example, a pupil who thinks the largest prime factor of 418 is 0, 1, 418, or some other number not specifically listed, will choose "none of these" and receive credit.

Unless great care is exercised, use of the "none-of-these" option may vitiate the major advantage of the multiple-choice item. In the usual multiple-choice item, minor flaws in the stem or in the responses are not likely to affect the validity of the item seriously because the intent of the item remains clear. Note what may happen, however, if the "none-of-these" option is introduced into the illustrative item on the boiling point of water:

At what temperature does water boil?
(1) 50°
(2) 73°
(3) 182°
(4) 212°
(5) none of these

The brighter pupils may once again read into the item more than was intended by the teacher and "none of these" may become a defensible answer.

Certainly, when the alternatives are numerical approximations the "none-of-these" response is an open invitation to trouble. For example:

The value of π is
(1) 0 (2) 2.17 (3) 3 (4) $\frac{22}{7}$ (5) none of these

A pupil might answer "none of these" on the basis that $\frac{22}{7}$ is not the *exact* value of π. This difficulty would not arise were it not for the presence of "none of these" as a response.

E. MATCHING ITEMS

1. Group only homogeneous premises and homogeneous responses in a single matching item.

The chief problem in the construction of matching exercises is to find homogeneous premises and homogeneous responses for which a meaningful basis for matching exists. When the lists are not homogeneous, the item measures only superficial verbal associations and is easily solved by those who have only vague concepts. In the illustrative example below, the items are so heterogeneous that each item in the first column is very obviously related to one and only one item in the second column. The others constitute totally implausible distractors.

Column A	*Column B*
_____ a. member of NATO	(1) Winston Churchill
_____ b. choose the President	(2) France
_____ c. woman suffrage	(3) electoral college
_____ d. Prime Minister of England	(4) 6 years
_____ e. term of U. S. Senator	(5) 19th Amendment

2. Keep the lists of premises and responses relatively short.

The matching lists should be kept short so that they can be easily handled by the pupil. Some matching questions in school examinations have as many as 20 or 30 items in each list. If a pupil is to compare each of 30 premises with each of 30 responses, he is required to make a total of 30×30 or 900 comparisons. Obviously, this is too long and tedious a task to impose upon the pupil. Economy of testing time, generally a strong advantage of the matching question, is no longer achieved.

There are other disadvantages as well. The strong possibility that the pupil may overlook the correct answer, buried as it is in a mass of other responses, tends to reduce validity. From the teacher's point of view, it is difficult to maintain homogeneity in a long list of items.

Further, a long list of matching items usually means excessive concentration on one aspect of achievement, preventing proper distribution of emphasis on other aspects.

A popular size in teacher examinations is the 10-item matching unit, which seems to work fairly well. Probably the optimum size from the testing viewpoint, however, would be the matching question containing 5 to 8 items.

3. Arrange the lists of premises and responses for maximum clarity and convenience to the pupil.

The pupil's usual method of approach to a matching question is to read the first item in the left-hand column and then to scan the entire right-hand column for the matching response. Thus, each item in the left-hand column may be read only once, while each response in the right-hand column may be considered and reconsidered many times. The matching item should be arranged in the most convenient form for the pupil. The longer, more complex statements should be used as premises and placed in the left-hand column, so that each need be read only once. The shorter, simpler statements, such as names, dates or brief phrases, should be used as responses and placed in the right-hand column. Short statements can be read more quickly, kept in mind more easily and compared more readily. In addition, if at all possible, the list of responses should be arranged in some logical order to simplify further the task of matching. For example, dates can be arranged in chronological order. Names can be alphabetized. The simpler the task posed for the pupil by a matching item, the more effective the item is likely to be.

4. Explain clearly in the directions the basis upon which items are to be matched and the procedure to be followed.

In simple matching exercises the basis for matching may be almost self-evident, but it is generally a good procedure to be as explicit as possible in the directions. Advising the pupil in the directions that, for example, musical selections are to be matched with their composers, mechanical tools with their functions or political events with the Presidents in whose administrations they occurred, serves to give the pupil an initial orientation that is very helpful to him. In the illustrative item below, there is likely to be considerable fumbling and confusion on the part of the pupils. Probably some pupils will never be able to work out a proper basis for classification on their own. The item is considerably improved by making the basis of classification explicit in the directions.

(Poor) On the line at the left of each item in Column A write the number of the matching item in Column B.

Column A	*Column B*
_____ a. lens	1. barometer
_____ b. mercury	2. electric light bulb
_____ c. vacuum tube	3. gasoline engine
_____ d. electromagnet	4. microscope
_____ e. filament	5. periscope
	6. radio
	7. telephone

(Improved) On the line at the left of each item in Column A write the number of the device in Column B in which it is used.

It is also important that, wherever a matching question varies from the usual form in which each premise is to be matched with one and only one response, the pupil be so informed. In some cases the teacher may wish to have the pupil indicate more than one response for a single premise, or the same response for two or more premises. Unless the pupil is aware that such answers are permissible, or even expected, he is placed under a severe handicap.

5. Avoid extraneous clues.

Any aspect of a matching question that provides an extraneous basis for correctly associating a premise and a response constitutes an undesirable clue. As in the case of multiple-choice items, such clues may assume a variety of forms—rote verbal associations, repetition of keywords, grammatical interrelationships in number or person etc. Certainly, when the premise calls for a Chinese general and the list of responses contains only one Chinese name, the probability of guessing the right answer approaches certainty. To the extent that pupils' answers reflect extraneous clues rather than competence in the field of study, the validity of the test as a measure of such competence is reduced.

6. Provide for extra responses to reduce guessing.

One of the advantages of the matching question is that it tends to minimize the guessing factor. In a 10-item matching question, for example, if the pupil does not know the answer to one of the items, his chances of guessing the right response are rather small. In many teacher examinations, however, the practice is to offer a list of premises and a list of responses of equal size. As a result, if a pupil knows 8 items of the 10, his chance of guessing correctly the responses to each of the remaining 2 items becomes 1 out of 2. If he knows 9 of

the 10, he can answer the remaining item correctly merely by a process of elimination. It is therefore generally considered good practice to have the number of responses exceed the number of premises. If there are 5 items, 7 responses might be offered. If there are 7 items, 10 responses might be offered.

IV. The Mechanical Features of Classroom Tests

It is not unusual to find that considerable pains have been taken in drawing up valid and effective test items, but that little or no attention has been given to the manner in which the items are put together as a test. The mechanical features of a test—how the test is printed and arranged, what instructions are given to the candidates, how the test is to be scored—are no less important than the test items themselves. Some mechanical features have a definite bearing on pupil performance and therefore influence the significance of the test scores. Other features have little or no bearing upon pupil performance but are designed to facilitate the handling and scoring of papers by the teacher. Careful attention by the teacher to test mechanics will not only enhance the value of the test as an evaluation tool but will pay handsome dividends in saving teacher time and effort.

A. ITEM FORMAT

All items should be numbered. Teachers sometimes omit item numbers in classroom tests because they feel no particular need for item numbers at the moment. However, some means of identifying individual items is nearly always found desirable afterward.

Each item should be presented in its entirety on the page. The pupil should not have to turn a page in the middle of a question. Similarly, any reference material (tables, graphs, reading selections etc.) should be on the same page as the item referring to it, or at least on a facing page. Breaking up item units may introduce unnecessary hazards and difficulties for the pupils. Usually the test can be rearranged to avoid this.

With respect to the arrangement of the item itself on the printed page, no special problem is presented to the teacher by completion, true-false or matching items. These item styles have been more or less standardized. In the case of multiple-choice items, however, a number of possibilities are open to the teacher. When space is an important consideration, the multiple-choice item can be presented in compact style. This style permits the use of practically all the space on the page with a minimum of waste:

An example of an implied power is the power to (1) levy income taxes (2) issue patents (3) declare war (4) regulate wages in interstate industries

When the alternatives are long or complex, however, this compact style does not permit the individual choices to be readily differentiated. It becomes difficult for the pupil to consider each alternative as a whole, and to compare and contrast the choices, especially when two or more choices have some words in common. In such instances greater clarity can be obtained by listing the alternatives in a vertical column:

A feature of the Federal Constitution that was derived from colonial American practice rather than from English practice was the

(1) right of trial by jury
(2) right of the lower house to originate revenue bills
(3) principle of a two-house legislature
(4) principle of the separation of powers in a written Constitution

Very often a good compromise between the space saving feature of the compact style and the greater legibility of the vertical arrangement of alternatives is the following:

Strong antislavery sentiment was expressed in his writings by

(1) Bret Harte **(3) Washington Irving**
(2) Mark Twain **(4) William Lloyd Garrison**

B. SCORING ARRANGEMENTS

Teachers preparing objective classroom tests should consider the possible advantages of a special answer sheet. In a brief one-page test, using a special answer sheet has no great advantage over having the pupils indicate their answers on the question paper. In a 3- or 4-page test, however, if answers are written on the question paper, considerable scoring time is involved merely in handling the different scoring stencils, turning the pages and totaling the separate scores. If a separate answer sheet is used, the teacher need usually prepare only one scoring stencil and can score each pupil's answers quickly and accurately in a single operation. Use of a separate answer sheet also makes it possible, if the teacher wishes, to use the same question papers over and over again.

If answers to objective items are to be indicated on the question paper, special consideration should be given to convenience and accuracy of scoring. The general aims should be to require as little time as possible for the pupil to designate his answer, to have the answers

placed in convenient location for scoring and to have the answers indicated in a manner that will help avoid possible misinterpretations.

To save time, it is advisable to have the pupil mark or designate the answer rather than write it out in full. In some tests, particularly those for younger children, the pupil is directed to underline, check or circle the correct answer. However, such responses are not convenient to score. More frequently the pupil is directed to indicate the correct answer by a code—a number or letter identifying the response. Research indicates that pupils can perform this translation of answer to symbol accurately and quickly, with no significant loss in validity.

To facilitate scoring, the pupils should be directed to mark their answers in spaces arranged in a vertical column. (A zig-zag, irregular pattern of responses across the page is time-consuming for the scorer and conducive to scoring errors.) Some teachers prefer to have the answers written on the left-hand side of the page, so that each answer is located immediately before the corresponding item number. Other teachers find it more convenient when the answers are written at the right-hand side of the page. In that case, it is desirable to make clear which answer space belongs to each item by using guidelines from the items to the answer spaces, or by repeating the item numbers next to the appropriate answer spaces.

Experience indicates that certain scoring symbols lead to scoring problems. Thus, "T" and "F" are often difficult to distinguish in careless penmanship. Pupils will sometimes take advantage of this weakness by deliberately hedging and writing an answer that can be interpreted either way. One possible solution is to use the symbol "+" for true and "O" for false. Another is to print the letter T and F beside each item and have the pupil circle the correct answer. A similar problem occurs in the multiple-choice item, where the letters *a*, *b*, *c* and *d* may be confused if carelessly written. For this reason many teachers use the capital letters A, B, C and D or arabic numerals 1, 2, 3 and 4 to designate alternatives.

C. DISTRIBUTION OF CORRECT RESPONSES

It is generally considered good practice in objective tests to avoid having too many consecutive items with the same answer. Unless the teacher checks the distribution of key answers, there may happen to be 6 or 8 "true" or "B" answers in a row. Some pupils who discover this are likely to go along with the trend quite readily, and even guess that the next item will continue the trend. Others may feel that they must be wrong and will be tempted to go back and reconsider some of

the questions. Thus, extraneous personality factors may be injected into the test score.

For the same reason, care is usually taken that no single answer appears with disproportionate frequency in the test as a whole. True and false answers are usually distributed with about equal frequency in the test as a whole, as are A, B, C, D answers in a multiple-choice test. Another reason for this precaution is to prevent the pupil who goes through a test and arbitrarily puts down the same answer for all items, or for all guessed items, from getting a high score. An evenly balanced distribution of key answers forestalls the accident of such a pupil hitting by chance on the answer which is usually correct.

Finally, patterning of answers should be avoided. Some teachers, to make scoring easy, may be tempted to set up a repeating TFFFT pattern of answers, or a repeating CADB pattern. The danger is that the pattern may become apparent to the pupil.

D. GROUPING AND ARRANGEMENT OF ITEMS

Should the questions in an examination be grouped according to content or according to question type? Most teachers will place the essay questions in a separate group, almost always at the end of the test. In the objective portion, however, practices vary. Some teachers will group all the multiple-choice items together, regardless of content. Some will follow a strict subject arrangement, so that 10 true-false items may be followed in order by 15 multiple-choice, 5 matching, 10 true-false, 5 multiple-choice etc.

It is generally agreed that, wherever possible, it is preferable to arrange the items in a test homogeneously by item type. This procedure is based upon the assumption that it is generally less confusing for the pupil to shift from one content area to another than from one item type to another. In addition, the test itself is simplified because it is unnecessary to repeat the same special directions throughout the test each time an item type reappears. Where indiscriminate mixing of subject matter is likely to be confusing, as in mixing vocabulary and spelling items, or in mixing questions about widely different historical periods, the items may be grouped by content *within* each item type.

Similarly, it is generally agreed that the items in a test should be arranged in order of difficulty rather than on the basis of chronology, order of teaching or some other aspect of content. A test should begin with the easier items and progress gradually to the more difficult ones. In this way the pupil enjoys a warmup period on material he can handle readily. He is not likely to be discouraged from doing his best

by early frustration in the test. And he is less likely to spend undue time puzzling over difficult questions while he leaves easy questions unanswered. When a test is set up in separate parts, either on the basis of item type or item content, the items within each part may be arranged in order of difficulty.

E. DESIGNATING CREDIT ALLOWANCES

In developing the blueprint for a test the teacher determines the number of credits to be assigned to each part of the test and to each item. These credit allowances should be indicated on the examination paper. It may not be a matter of crucial importance to the pupil whether, for example, the true-false items are worth ½ point each and the multiple-choice items 1 point each. Presumably the pupil will do his best in any case. However, the credits assigned to different parts of the test will undoubtedly be of some help to the pupil in planning his time for answering the questions and reviewing the paper. It avoids the complaint, "I didn't know this part counted for more than half of the test or I would have spent more time on it!"

In the case of essay questions, credit designations become essential. Designating credit allowances in essay questions indicates to the pupil the relative emphasis to be given to the different parts of a question and suggests the desired scope or depth of treatment. In the illustrative question below, the answer to be expected from the pupil would be quite different if the credits for parts *a* and *b* were reversed, or if the whole question were to count 5 credits instead of 15.

a. Describe the circumstances under which each of *two* "third parties" has appeared in the political development of our country. (10)

b. Show how one of these parties had an important effect on our history. (5)

F. DIRECTIONS FOR ANSWERING QUESTIONS

Most pupils in our schools have had experience with objective test questions. Probably if a set of completion or multiple-choice questions were placed before the average pupil without any directions whatsoever, he would proceed quite naturally to answer the questions as expected—in somewhat the same routine manner in which a veteran milk wagon horse completes his familiar rounds. This may explain why carelessly drawn instructions in examinations do not cause even greater confusion than they apparently do.

The safest procedure for the teacher to follow is to assume that the pupil knows nothing about objective tests. Each set of items should

be introduced by a set of simple, clear, specific instructions. Often it is helpful to the pupil to highlight the instructions by variation in type or margins. In the case of new or unusual item types, a few samples or illustrative questions and answers at appropriate points in the test will go far toward clarifying the directions.

G. CORRECTION FOR GUESSING

What to do about guessing and chance success is a question that has plagued test makers from the very beginning of objective testing. In fact, complete agreement on this point has not yet been attained.

Guessing can enter into objective test items to various degrees, depending upon the item type. A general formula, applicable to all item types, can be applied to correct for guessing. The theory behind this formula is that guessing enters into all wrong answers on an entirely random basis. This is a highly questionable assumption. As every teacher knows, guesses by different pupils on different items are not entirely random but are based on different amounts of information. Many guesses, indeed, are really based on shrewd analysis. Further, when pupils are instructed not to guess, their behavior will be affected by personality differences. The cautious individual will refuse to answer unless absolutely certain of the answer, while the more daring soul will answer quite readily when he has only a smattering of information. No uniform formula can correct for human variations.

In the case of teacher-made tests designed for local use and liberally timed, it is doubtful whether the use of a correction formula is worth the trouble. Unless some of the pupils omit a disproportionately large number of items, the correlation between the corrected and uncorrected scores will usually be quite high. The teacher may therefore be fairly certain that the pupils will be ranged in approximately the same order whether the scores are corrected for guessing or not.

In view of these considerations, it is suggested that teachers need not correct for guessing in the ordinary classroom test. Pupils should be directed to try all questions. When they are not sure of the answer, they should be advised to give the best answer they can. The hope is that in this way the guessing factor will be kept as uniform as possible for all pupils.

Whatever decision the teacher reaches concerning correction for guessing, the pupil should be informed of the rules of the game. If deductions are to be made for wrong answers, the pupils should be so advised and cautioned not to guess. If no deductions are to be made

for wrong answers, the pupils should be advised to answer every question according to their best judgment.

H. ALLOWING CHOICE OF ITEMS

The usual practice in classroom essay tests is to allow pupils a choice of questions to answer, such as 4 out of 7, or 6 out of 8. Within the essay question itself, there is often a choice of sections or subdivisions that may be answered. Such provision for choice of items is usually an attempt on the part of a teacher to minimize possible accidents of content sampling. Where a final examination has only 4 or 5 essay questions, for example, large blocks of the curriculum cannot be tested. If the questions in the examination were changed so that areas omitted were substituted for those tested, the scores of the pupils might change drastically because of their individual strong points and weak points. Allowing the pupil some choice will tend to reduce such possible unfairness due to sampling. Of course, it is most important that the directions indicating the choice be prominent, clear and explicit. When some pupil answers too many questions or not enough questions, it is always a difficult problem for the teacher to handle the situation equitably.

What is not commonly realized is that the scores of two pupils in a test where choice is allowed are not strictly comparable unless they have made the same choices—no more than if they had taken different tests. In truth, they have taken different tests. As many different tests are involved as there are different possible combinations of choices. Care must therefore be taken that the choice is not so wide, and that the questions among which choice is to be exercised do not differ so markedly in difficulty, that comparability of pupil scores is likely to be seriously affected. Indeed, where strict comparability of scores is desirable, or where no pupil is likely to be handicapped unfairly by the content sampling of the test, then all pupils might well be required to answer the same questions. In objective-type tests, because of the wider range of sampling, there is ordinarily little need to provide for choice. If choice is allowed in objective tests, it is desirable that the teacher avoid imposing upon the pupil a clerical counting job that will be a potential source of trouble. Some tests, for example, require that the pupil answer 70 true-false statements out of 100. Inevitably the pupil is going to spend an inordinate length of time counting and recounting the 30 items he will omit, or more likely the 70 he will answer. Just as inevitably, many pupils will answer 69 items or 71 items. The pupil's task would be greatly simplified if the 100 items were arranged, for example, in 10 groups of 10 items each, with

a choice of 7 in each group. In addition, this type of procedure allows the teacher to exercise greater control over the sampling of the test. If a pupil may choose any 70 items in a test, he may be able to avoid entirely some important areas of the curriculum, even though they are covered in the test. If a choice is exercised only within smaller groups, each directed at a different content area, the items selected by the pupil must necessarily cover the entire range and scope of the test.

I. PRINTING AND DUPLICATING

An examination consisting of only a few questions can be administered, if the teacher wishes, by dictating the questions or by writing them on the blackboard. This procedure, however, is inadvisable even for brief tests and impossible for longer ones. The best procedure is to reproduce sufficient copies to supply each pupil with his own individual copy of the test.

The reproduction of a test should, of course, follow the same high standards of any printed material. The type size should be sufficiently large, adequate space should be left between lines, and the printing should be clear. Classroom tests are especially likely to suffer from confusing misprints and need to be very carefully proofread. Consideration must also be given to the optimum length of line for the type of material in the test and the grade level of the pupils. Very long lines or very short lines tend to interfere with comprehension. A good practical guide is that the length of line in a test might well approximate the length of line in the pupils' standard reading materials.

V. Checklist for Reviewing Local School Tests

This checklist covers the major test construction principles which teachers should observe in the preparation of local school tests. A negative answer to any question in the checklist indicates variation of local practice from accepted principles of test construction, suggesting that a reevaluation of local practice may be desirable with a view toward possible improvement of the test.

The items in the checklist cannot be added like the items of an examination to provide a "score" for a local school test. The questions are not of equal importance. Though a test may be satisfactory in all respects but one, that single deficiency may be sufficient to impair seriously the value of the test as a whole. The purpose of the checklist, rather, is diagnostic. It is designed to help insure that in the development of a local school test no important factor likely to influence the effectiveness of the test will be overlooked.

For a more detailed explanation of any item in the checklist, the teacher is referred to the section of this bulletin in which that point is discussed.

Classroom Test Checklist

I. The Test Items

A. Essay Questions

1. Are essay questions restricted to measurement of objectives not readily measured by other item types?
2. Are essay questions framed around specific problems, adequately delimited in scope?
3. In general, is use made of a large number of brief essays rather than one or two extended essays?
4. Does each essay question indicate clearly and accurately the desired extent and depth of the answer?
   a. Does the question indicate clearly how many reasons, examples, arguments etc. are expected for full credit?
   b. Do the directions to "explain," "outline," "state," "compare" etc. indicate accurately the type of answer that will receive full credit?

B. Completion Items

1. Is each statement sufficiently unambiguous to limit the correct answer to one or two specific words or phrases?
2. Is an excessive number of blank spaces avoided?

3. Does the omitted part of each question come at the end or near the end of the statement?
4. Are the items free from extraneous clues due to grammatical structure, length of blank space etc.?
5. Is use of the completion form generally avoided when items are essentially true-false in nature?
6. Do computational problems indicate the expected degree of precision in the answer?
   a. Does the question indicate clearly the extent to which approximations or fractional answers are to be rounded?
   b. Does the question indicate clearly whether units (such as square yards or feet per second) are to be included by the pupil in his answer?

C. True-False Items

1. Is each item definite and unambiguous in meaning?
2. Are the items based upon statements that are absolutely true or false, without qualifications or exceptions?
3. Has the central point of each question been highlighted by placing it in a prominent position?
4. Are the items free from double-barreled statements that are partly true and partly false?
5. Are the items expressed as simply and directly as possible, without many qualifying clauses?
6. Are trick questions avoided?
7. Is excessive "window dressing" in the items avoided?
8. Are negative questions avoided so far as possible?
9. Are the items free from specific determiners such as "always" and "usually"?
10. In the modified true-false type of item, is the word to be corrected clearly indicated by underlining or special type?

D. Multiple-Choice Items

1. Is use of the direct question or incomplete statement form consistent with the most effective presentation of the individual items?
2. Are the items presented in clear and simple language, with vocabulary kept as simple as possible?

3. Does each item have one and only one correct answer?
4. Is each item concerned with a single central problem?
5. Is the central problem of each item stated clearly and completely in the stem?
6. Does the stem, so far as possible, include all words repeated in the responses?
7. Are negative statements avoided?
8. Is excessive "window dressing" avoided?
9. Do the responses or choices come at the ends of the incomplete statements?
10. Are the responses grammatically consistent with the stem and parallel with one another in form?
11. Are all responses plausible and attractive to pupils who lack the information or ability tested by the item?
12. Are the responses, so far as possible, arranged in numerical or logical order?
13. Are the responses independent and mutually exclusive?
14. Are the items free from extraneous clues due to grammatical inconsistencies, rote verbal associations, length of response etc.?
15. Is the "none-of-these" option used only when appropriate?

E. Matching Items

1. Do the premises and responses constitute homogeneous lists, each grouped around a single concept?
2. Are the lists of premises and responses relatively short?
3. Are the matching lists conveniently arranged?
   a. Are the longer, more complex statements used as premises and the shorter statements as responses?
   b. Wherever possible, are the responses arranged in some logical order to simplify matching?
4. Do the directions indicate clearly the basis upon which the lists are to be matched?
   a. If a single premise is to be matched with several responses, is the pupil advised that this is permissible?
   b. If a single response is to be matched with several premises, is the pupil advised that this is permissible?
5. Are the matching lists free from extraneous clues due to grammatical construction, rote verbal associations etc.?
6. Is the list of responses longer than the list of premises to preclude guessing by elimination?

II. Mechanical Features of the Test

A. Item Format

1. Are the items in the test numbered?
2. Is each item complete on a page?
3. Does the reference material for an item appear on the same page as the item or on a facing page?
4. Are the item responses arranged to achieve both legibility and economy of space?

B. Scoring Arrangements

1. Has consideration been given to the practicability of a separate answer sheet?
2. Are answers to be indicated by symbols rather than underlining or copying?
3. Are answer spaces placed in a vertical column for easy scoring?
4. If answer spaces are placed at the right of the page, is each answer space clearly associated with its corresponding item?
5. Are the answer symbols to be used by the pupils free from possible ambiguity due to careless penmanship or deliberate hedging?
6. Are the answer symbols to be used by the pupils free from confusion with the substance or content of the responses?

C. Distribution of Correct Responses

1. Are correct answers distributed so that the same answer does not appear in a long series of consecutive questions?
2. Are correct answers distributed to avoid an excessive proportion of items in the test with the same answer?
3. Is patterning of answers in a fixed repeating sequence avoided?

D. Grouping and Arrangement of Items

1. Are items of the same type requiring the same directions grouped together in the test?
2. Where juxtaposition of items of markedly dissimilar content is likely to cause confusion, are items grouped by content within each item type grouping?
3. Are items generally arranged from easy to more difficult within the test as a whole and within each major subdivision of the test?

E. Designating Credit Allowances

1. Are credits indicated for the major sections of the test?
2. Is the credit allowance for each item clear to the pupil?
3. Where questions have subdivisions, especially in essay questions, are credits indicated for each of the parts of the question?

F. Directions for Answering Questions

1. Are simple, clear and specific directions given for each different item type in the test?
2. Are directions clearly set off from the rest of the test by appropriate spacing or type style?
3. Is effective use made of sample questions and answers to help clarify directions for unusual item types?

G. Correction for Guessing

1. If deductions are to be made for wrong answers, are pupils instructed not to guess?
2. If no deductions are to be made for wrong answers, are pupils advised to answer every question according to their best judgment?

H. Allowing Choice of Items

1. Is the degree of choice sufficiently limited, and are the questions among which choice is allowed sufficiently similar in difficulty, to maintain reasonable comparability of pupils' scores?
2. Are the directions covering choice prominent, clear and explicit?
3. Is choice exercised within relatively small groups of items rather than among many items?

I. Printing and Duplicating

1. Has the test been duplicated to provide individual pupil copies?
2. Is the test free from annoying and confusing typographical errors?
3. Is the legibility of the test satisfactory from the viewpoint of type size, adequacy of spacing and clarity of printing?
4. Is the length of line neither too long nor too short for easy comprehension?

Bibliography

Adkins, Dorothy C. Construction and analysis of achievement tests. U. S. Government Printing Office, Washington 25, D. C. 1947

Greene, Harry A., Jorgensen, Albert N. & Gerberich, J. Raymond. Measurement and evaluation in the secondary school. N. Y. Longmans, Green and Co. 1943

Hawkes, Herbert E., Lindquist, E. F. & Mann, C. R. The construction and use of achievement examinations. N. Y. Houghton Mifflin Co. 1936

Lindquist, E. F. Educational measurement. American Council on Education. Washington, D. C. 1951

Remmers, H. S. & Gage, N. L. Educational measurement and evaluation. N. Y. Harper and Brothers. 1943

Ruch, G. M. The objective or new-type examination. Scott, Foresman and Co. 1929

Weitzman, Ellis, & McNamara, Walter J. Constructing classroom examinations. Chicago. Science Research Associates. 1949